Neighbor-Joining with Interval Methods

D. Levy[1], Raazesh Sainudiin[2], R. Yoshida[3], and L. Pachter[1]

[1]University of California at Berkeley, Berkeley, CA 94720, USA; [2]Cornell University, Ithaca NY 14850; [3]Duke University, Durham, NC 27708, USA

The software package `MJOIN` is available at `http://bio.math.berkeley.edu/mjoin/`

Introduction

- The Neighbor-Joining algorithm is a recursive procedure to reconstruct a phylogenetic tree using a transformation of pairwise distances between leaves for identifying cherries in the tree.
- Pachter and Speyer showed that we can recover an n-leaf tree from the weights of m-leaf subtrees if $n \geq 2m-1$ [PS04].
- We generalized the cherry picking criterion with estimates of the weights of m-leaf subtrees.
- We showed that a reconstructed tree from such weights is more accurate than one using pairwise distances.
- This leads to an improved neighbor-joining algorithm whose total running time is still polynomial in the number of taxa.

Neighbor Joining with Pairwise Distances

Theorem. (the cherry picking criterion) [SN87, SK88]
Suppose $D(ij)$ is a pairwise distance between taxa i and j. Then, $\{i,j\}$ is a cherry if $A_{ij} = D(ij) - (r_i + r_j)/(n-2)$, where $r_i := \sum_{k=1}^{n} D(ik)$, is minimal.

Idea. Initialize a star-like tree and find a cherry. Then we compute branch length from the interior node to each leaf. Repeat this process recursively until we find all cherries.

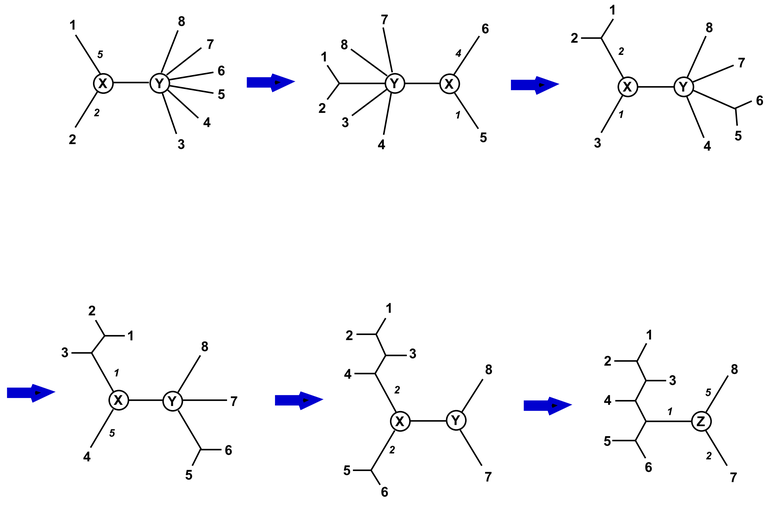

FIGURE 1: The traditional Neighbor Joining with pairwise distances.

Neighbor Joining with Subtree Weights

Notation. Let $[n]$ denote the set $\{1,2,...,n\}$ and $\binom{[n]}{m}$ denote the set of all m-element subsets of $[n]$.

Definition. A *m-dissimilarity map* is a function $D : \binom{[n]}{m} \to \mathbb{R}_{\geq 0}$. In terms of phylogeny, this corresponds to the weights of m-subtree weights of a tree T.

Theorem. Let D_m be be an m-dissimilarity map on n leaves, $D_m : \binom{[n]}{m} \to \mathbb{R}_{\geq 0}$ correspond to the weights of m-subtree weights of a tree T and we define

$$S(ij) := \sum_{X \in \binom{[n]\setminus\{i,j\}}{m-2}} D_m(ijX).$$

Then $S(ij)$ is a tree metric.
Furthermore, if T' is the additive tree corresponding to this tree metric then T' and T have the same tree topology and there is an invertible linear map between their edge weights.

Algorithm. (Neighbor Joining with Subtree Weights)

- **Input**: n many DNA sequences.
- **Output**: A phylogenetic tree T with n leaves.

1. Compute all m-subtree weights via the maximum likelihood.
2. Compute $S(ij)$ for each pair of leaves i and j.
3. Apply Neighbor Joining method with a tree metric $S(ij)$ and obtain additive tree T'.
4. Using a linear mapping, obtain a weight of each internal edge and each leaf edge of T.

Cherry Picking Theorem

Theorem. Let T be a tree with n leaves and no nodes of degree 2 and let m be an integer satisfying $2 \leq m \leq n-2$. Let $D : \binom{[n]}{m} \to \mathbb{R}_{\geq 0}$ be the m-dissimilarity map corresponding to the weights of the subtrees of size m in T. If $Q_D(ab)$ is a minimal element of the matrix

$$Q_D(ab) = \binom{n-2}{m-1} \sum_{X \in \binom{[n]\setminus\{i,j\}}{m-2}} D(ijX) - \sum_{X \in \binom{[n]\setminus\{i\}}{m-1}} D(iX) - \sum_{X \in \binom{[n]\setminus\{j\}}{m-1}} D(jX)$$

then $\{a, b\}$ is a cherry in the tree T.

Note. The theorem by Saitou-Nei and Studier-Keppler is a corollary from Cherry Picking Theorem.

Time Complexity

If $m \geq 3$, the time complexity of this algorithm is $O(n^m)$, where n is the number of leaves of T and if $m = 2$, then the time complexity of this algorithm is $O(n^3)$.
Note: The running time complexity of the algorithm is $O(n^3)$ for both $m = 2$ and $m = 3$.

Interval Methods

- In [LYP04], Dissimilarity maps are computed via `fastDNAml` which implements a gradient flow algorithm with floating-point arithmetic.
- Instead, apply the rigorously enclosed maximum likelihood estimations [Sai04].
- Dissimilarity maps computed via the rigorously enclosed MLEs are guaranteed to be enclosed. Thus, reconstructed trees via the generalized NJ method with these dissimilarity maps are more accurate.

Computational Results

- **Problem:** Find the NJ tree for 21 *S-locus receptor kinase* (SRK) sequences [SWY+05] involved in the self/nonself discriminating self-incompatibility system of the mustard family [Nas02].
- **Result:** Symmetric difference ($\boldsymbol{\Delta}$) between 10,000 trees sampled from the likelihood function via MCMC and the trees reconstructed by 5 methods.
  DNAml was used in two ways: DNAml(A) is a basic search with no global rearrangements, whereas DNAml(B) applies a broader search with global rearrangements and 100 jumbled inputs.

$\boldsymbol{\Delta}$	NRGNJ	fastDNAml	DNAml(A)	DNAml(B)	TrExML
0	0	0	2	3608	0
2	0	0	1	471	0
4	171	6	3619	5614	0
6	5687	5	463	294	5
8	4134	3987	5636	13	71
10	8	5720	269	0	3634
12	0	272	10	0	652
14	0	10	0	0	5631
16	0	0	0	0	7

References

[LYP04] D Levy, R Yoshida, and L Pachter. Neighbor joining with subtree weights. *preprint*, 2004.

[Nas02] JB Nasrallah. Recognition and rejection of self in plant reproduction. *Science*, 296:305–308, 2002.

[PS04] L. Pachter and D. Speyer. Reconstructing trees from subtree weights. *Applied Mathematics Letters*, 17:615 – 621, 2004.

[Sai04] R Sainudiin. Enclosing the maximum likelihood of the simplest DNA model evolving on fixed topologies: towards a rigorous framework for phylogenetic inference. Technical Report BU1653-M, Department of Biol. Stats. and Comp. Bio., Cornell University, 2004.

[SK88] J. A. Studier and K. J. Keppler. A note on the neighbor-joining method of saito and nei. *Mol. Biol. Evol.*, 5:729 – 731, 1988.

[SN87] N. Saitou and M. Nei. The neighbor joining method: a new method for reconstructing phylogenetic trees. 1987.

[SWY+05] R Sainudiin, SW Wong, K Yogeeswaran, J Nasrallah, Z Yang, and R Nielsen. Detecting site-specific physicochemical selective pressures: applications to the class-I HLA of the human major histocompatibility complex and the SRK of the plant sporophytic self-incompatibility system. *Journal of Molecular Evolution*, in press, 2005.